Analyse des compagnies de production cinématographique en Europe

Mise en pratique de l'ETL pour la visualisation de données provenant des données relationnelles

Veerasingam Abilash
Lontchi Taboua Freddy
Hajjare Moutaki

1

I. Table des matières

II. Objectifs et présentation du projet ___ *3*

A. A propos de la base de données ___ **3**

B. Objectifs du projet ___ **3**

C. Technologies utilisées ___ **3**

1. PostgreSQL ___ 3
2. Talend ___ 4
3. Tableau ___ 4

D. Schéma de la Base de Données ___ **4**

III. ETL ___ *7*

E. Conception ___ **7**

F. Extraction ___ **8**

G. Transformation des données ___ **9**

H. Load ___ **10**

4. Définition des flux ETL et charger les données dans le Data Warehouse. ___ 10

I. Construction du Data Warehouse ___ **12**

IV. Visualisation avec Tableau ___ *14*

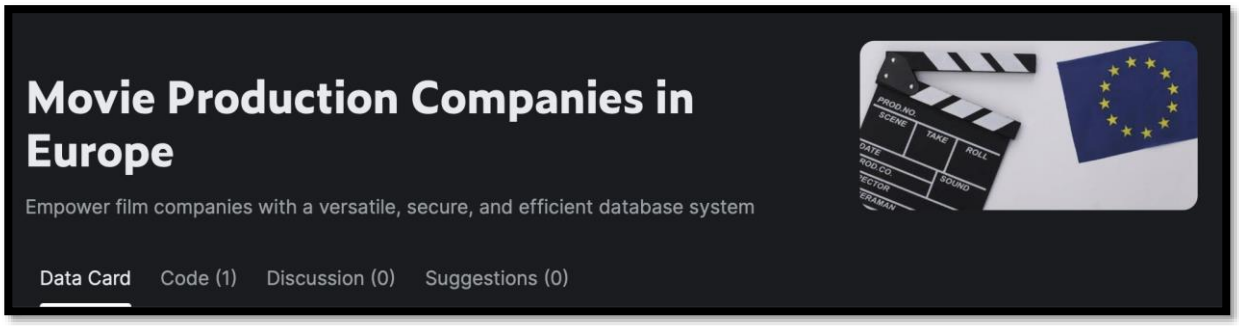
J. Connexion à la Base de Données ___ **14**

K. Création des visuels ___ **14**

5. Vue d'ensemble de l'organisation : ___ 14
6. Indicateurs KPI : ___ 16

V. Conclusion ___ *19*

II. Objectifs et présentation du projet



A. A propos de la base de données

L'ensemble de données analysé contient des enregistrements fictifs d'employés de sociétés de production cinématographique à travers l'Europe. L'ensemble de données est structuré en dix 20 fichiers, fournissant des informations détaillées sur les individus, notamment leur nom complet, leur sexe, leur date de naissance, leurs affiliations à l'entreprise, leurs rôles et leurs dates de début d'emploi. Ces données sont inestimables pour explorer la structure organisationnelle, la démographie de la main-d'œuvre et les rôles au sein de l'industrie européenne de la production cinématographique. Ces enregistrements aident les utilisateurs à naviguer et comprendre les capacités et la structure de la base de données dans un contexte théorique.

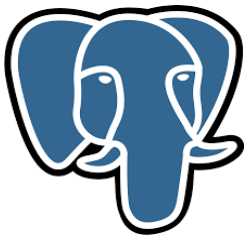
Source des données : kaggle

B. Objectifs du projet

Notre objectif est d'utiliser les techniques ETL sur Talend pour agréger les données afin de créer un nouveau schéma de données qui palliera le problème de lenteur crée par les jointures profondes des modelés relationnels lors de la visualisation. Notre data Warehouse sera basé sur ce nouveau schéma. Nous détaillerons les indicateurs visualisés plus bas.

C. Technologies utilisées

1. PostgreSQL



pgAdmin est un outil de gestion de base de données open source pour PostgreSQL, utilisé dans ce projet pour gérer la base de données des compagnies de production cinématographique en Europe. Il a permis de créer notre Datawarehouse et de stocker aussi une partie de nos données avant l'ETL.

2. Talend

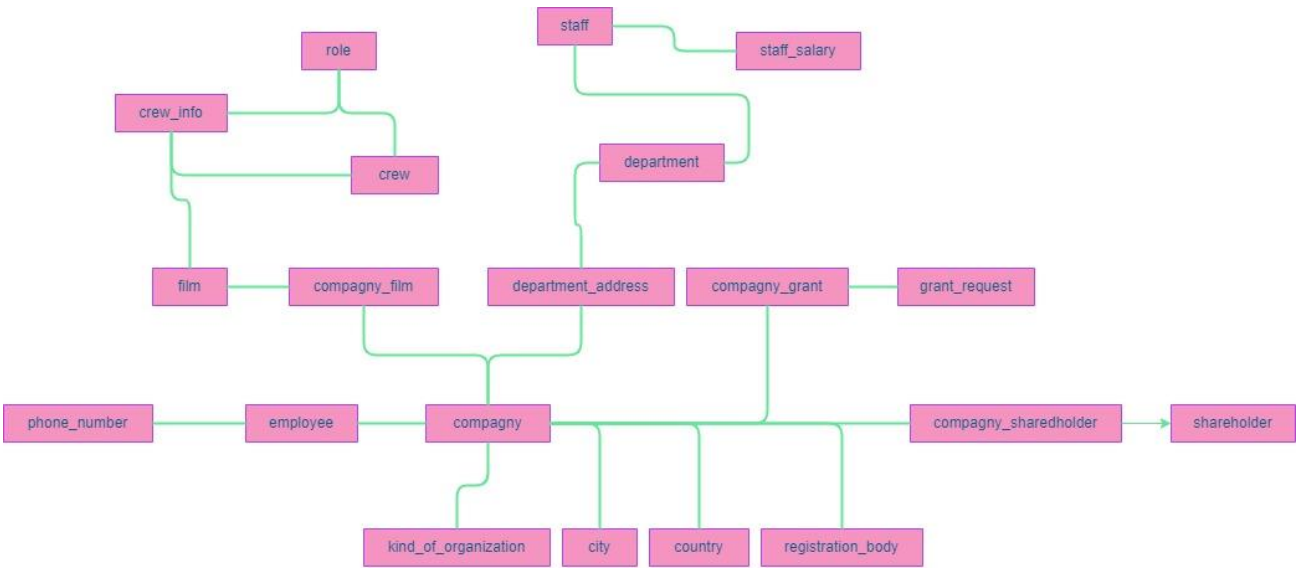
Talend est une plateforme d'intégration de données qui a été utilisée pour les processus ETL (Extract, Transform, Load) du dataset. Grâce à Talend, les données des compagnies de production cinématographique ont été collectées, une partie venant de PostgreSQL et une autre partie venant des fichiers plats csv. Puis, ces données ont été transformées et chargées dans le Data Warehouse sur PostgreSQL.

3. Tableau

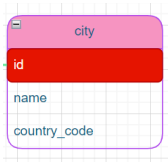
Tableau est un outil de visualisation de données utilisé pour créer des rapports interactifs à partir du dataset sur les compagnies de production cinématographique en Europe. Il a permis de visualiser les tendances et d'explorer les relations entre les différentes entités, offrant des insights précieux sur l'industrie cinématographique et facilitant la prise de décision stratégique.

D. Schéma de la Base de Données

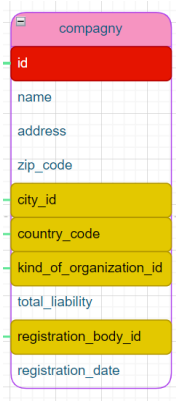
Cette photo par Auteur



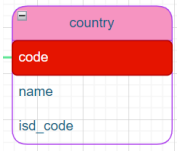
- **city.csv**



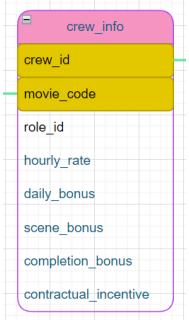
- **company.csv**



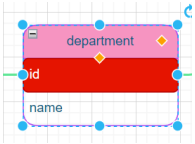
- **country.csv**



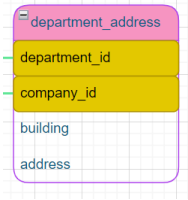
- **crew_info.csv**



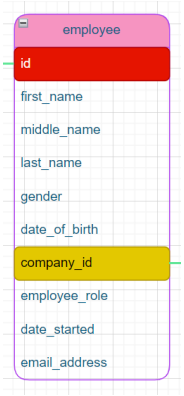
- **department.csv**



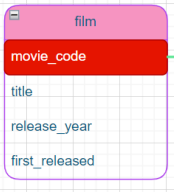
- **department_address.csv**



- **employee.csv**

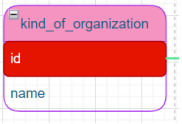


- **film.csv**

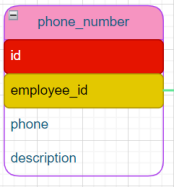


- **grant_request.csv**

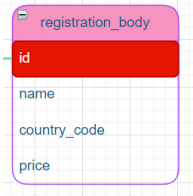
- **kind_of_organization.csv**



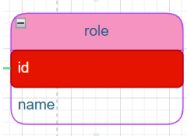
- **phone_number.csv**



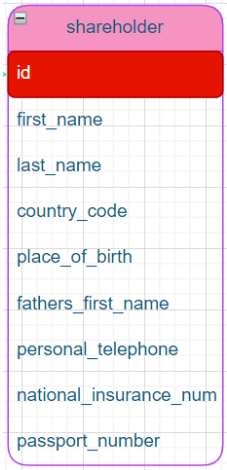
- **registration_body.csv**



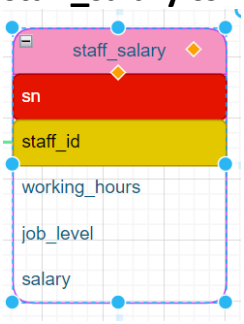
- **role.csv**



- **shareholder.csv**



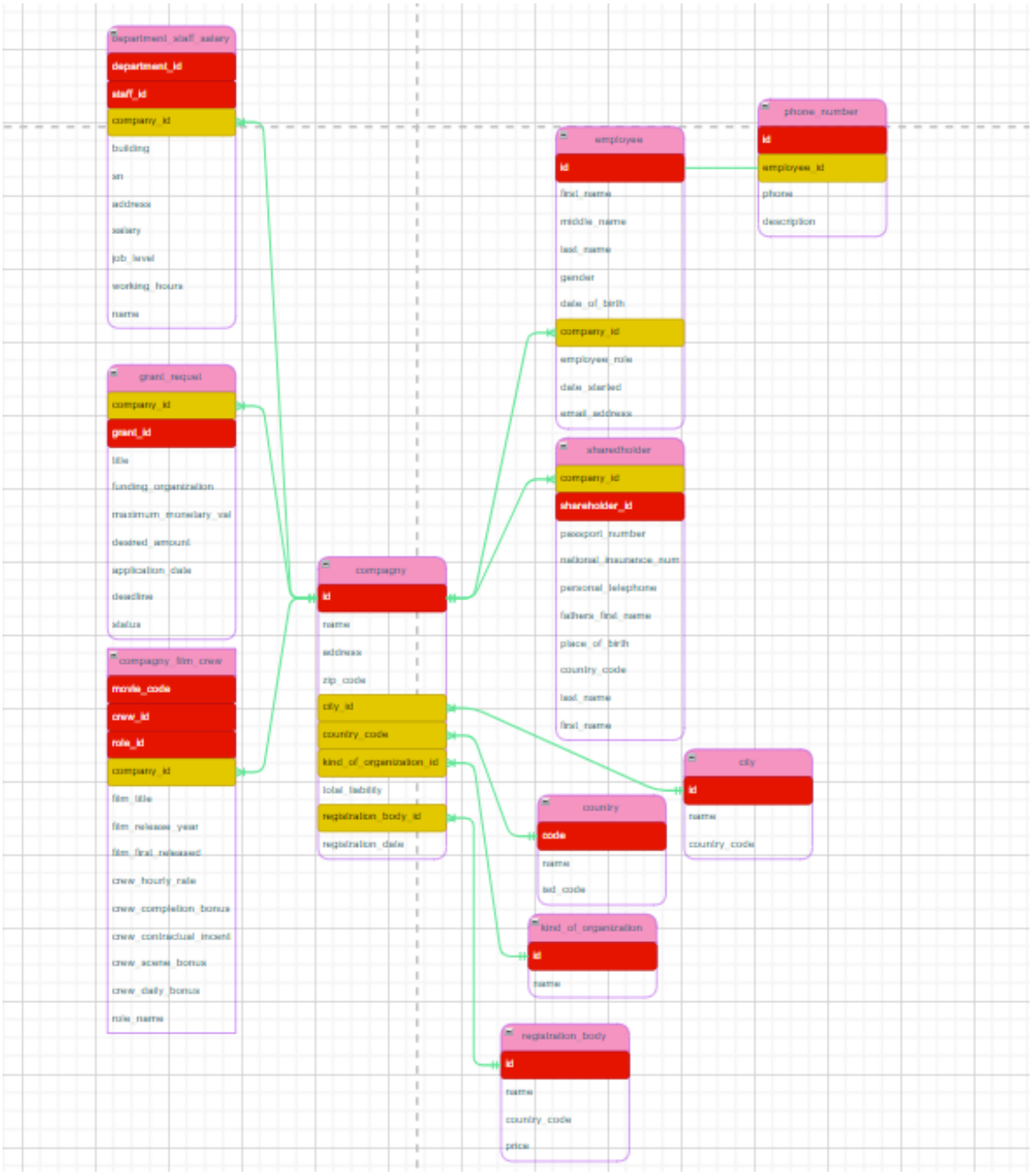
- **staff_salary.csv**



III. ETL

A. Conception

Nous avons regroupé les tables excentrées à plus d'un pas de la table company pour les reprocher le plus de celle-ci. Nous pouvons voir le résultat sur la figure ci-dessous.

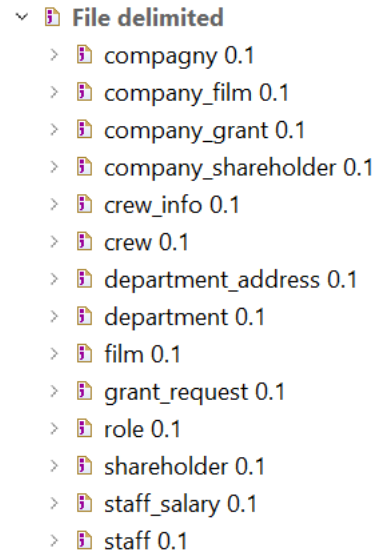


B. Extraction

Les données proviennent de deux sources :

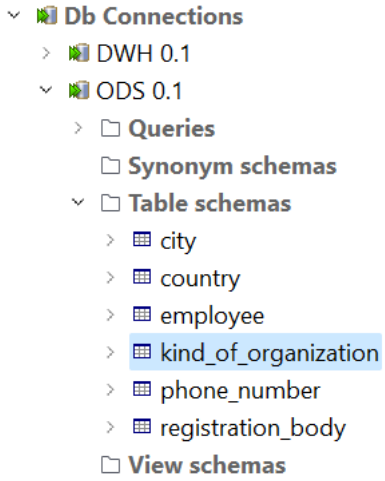
- **A partir des fichiers CSV**

Voici la liste des fichiers qui ont été importés.



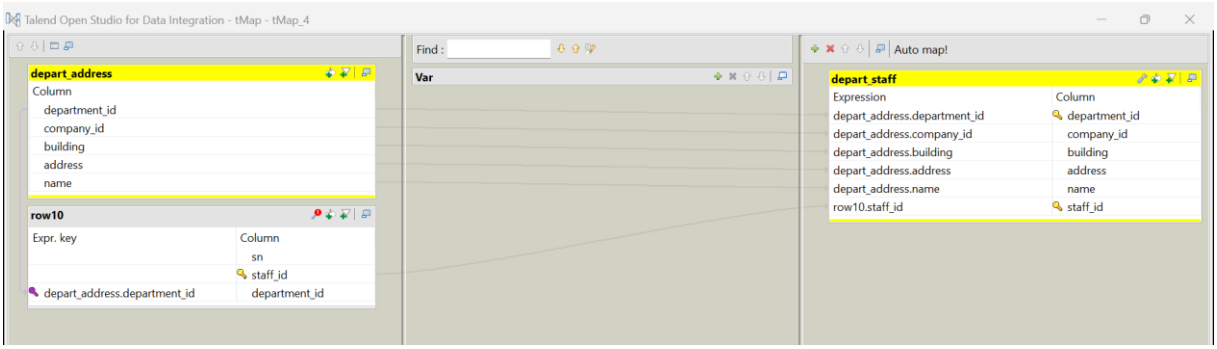
- **A partir de PostgreSQL**

  Une connexion a été établie vers le **schéma ODS** de notre base de données afin de récupérer le reste des fichiers.

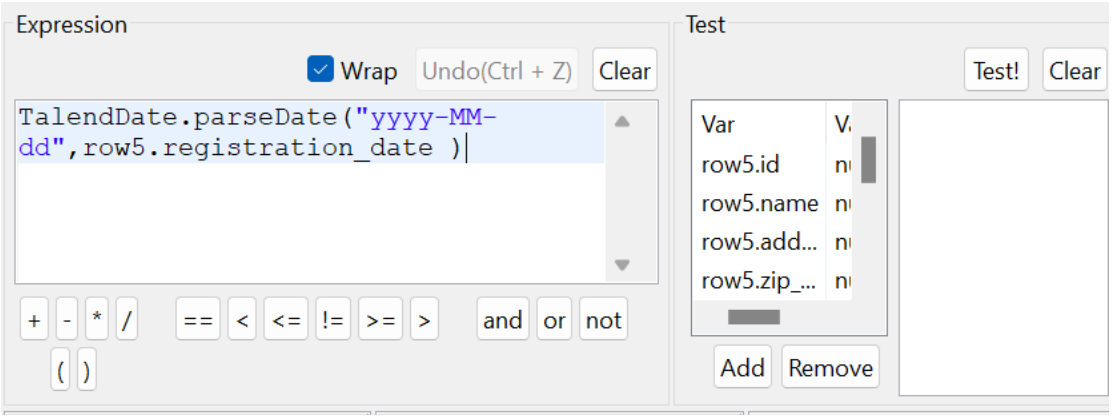


C. Transformation des données

Concernant la transformation des données, il s'agit principalement de l'agrégation des données grâce au composant **tmap** afin d'obtenir un model en étoile.

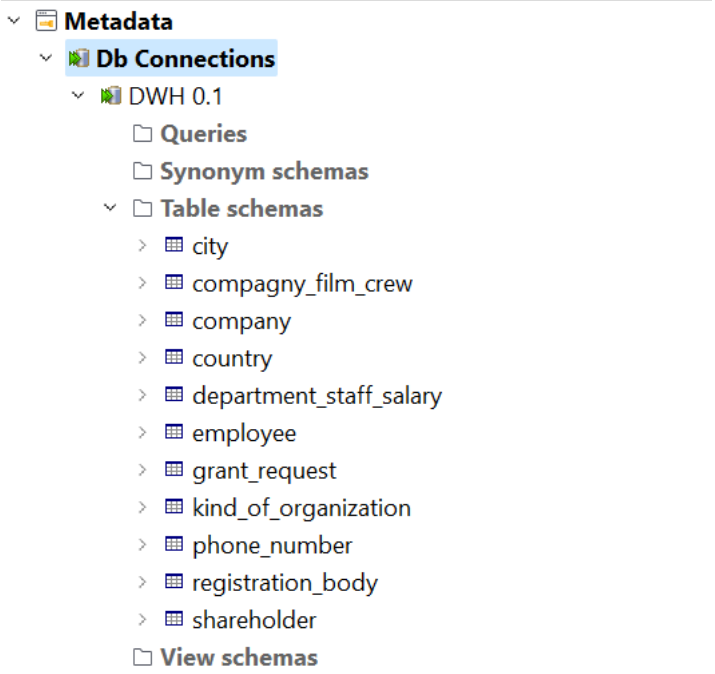


Nous avons aussi appliqué un formatage sur certaines dates qui n'avaient pas le bon format ou qui n'étaient pas reconnues comme des dates.



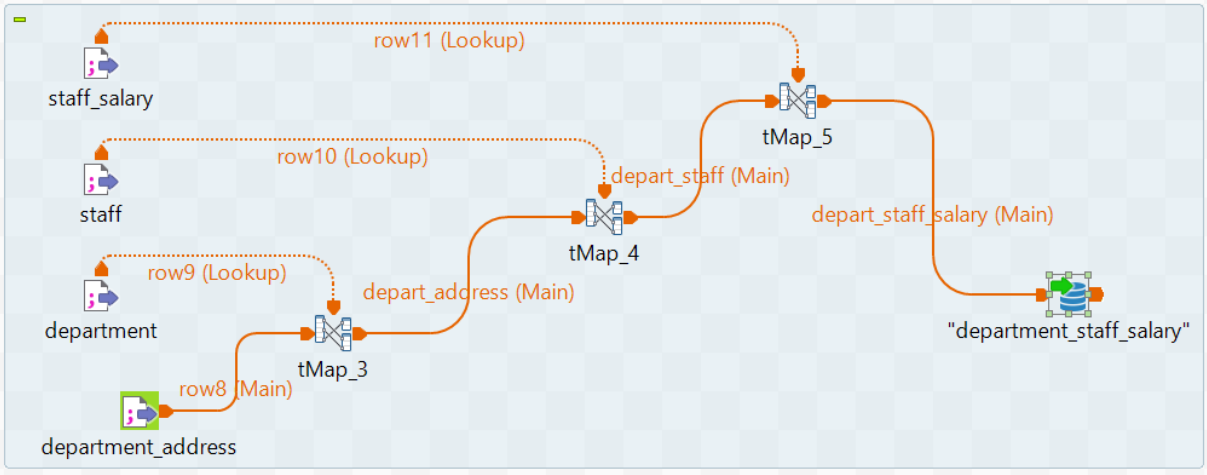
D. Load

Une deuxième connexion vers le a été faite vers le schéma **DWH** pour avoir accès au data Warehouse.

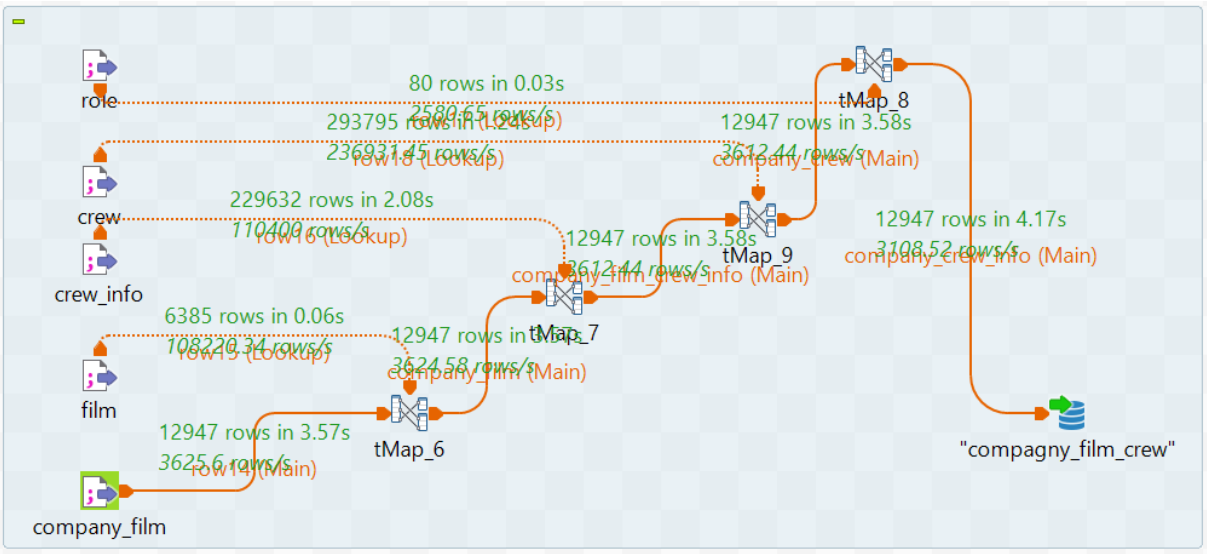


1. Définition des flux ETL et charger les données dans le Data Warehouse.

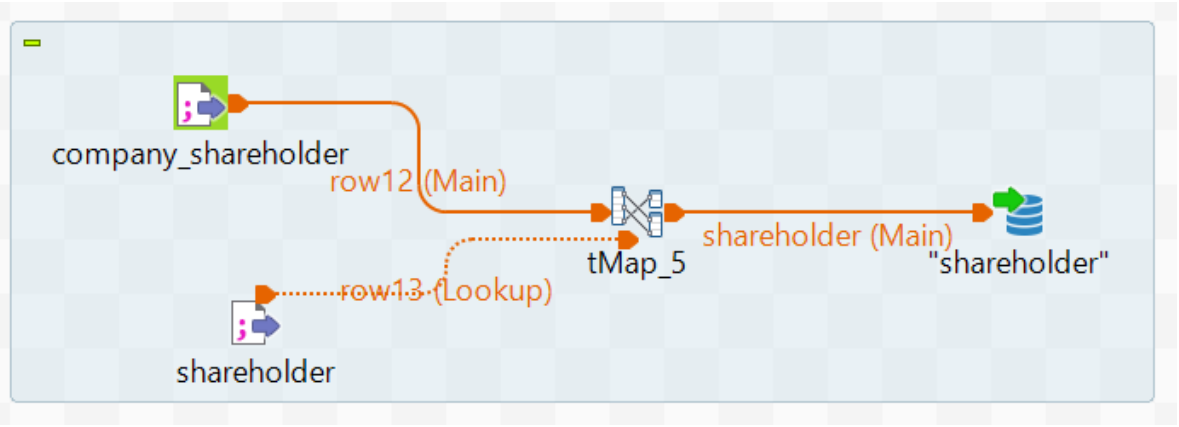
- Agrégation des tables (staff, staff_salary, departement et department adress) pour obtenir **department_staff_salary**



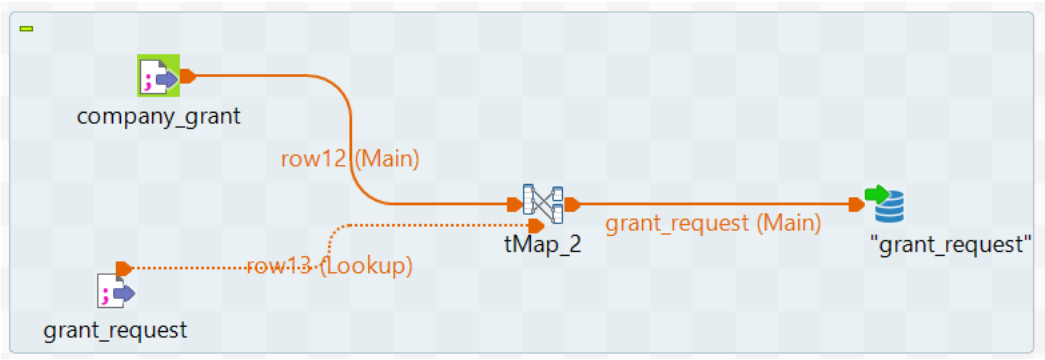
- Agrégation des tables (crew, role, crew_info, film, compagny) pour obtenir la table **compagny_film_crew**



- Agrégation des tables (compagny_sha et shareholder) pour obtenir la table **shareholder**

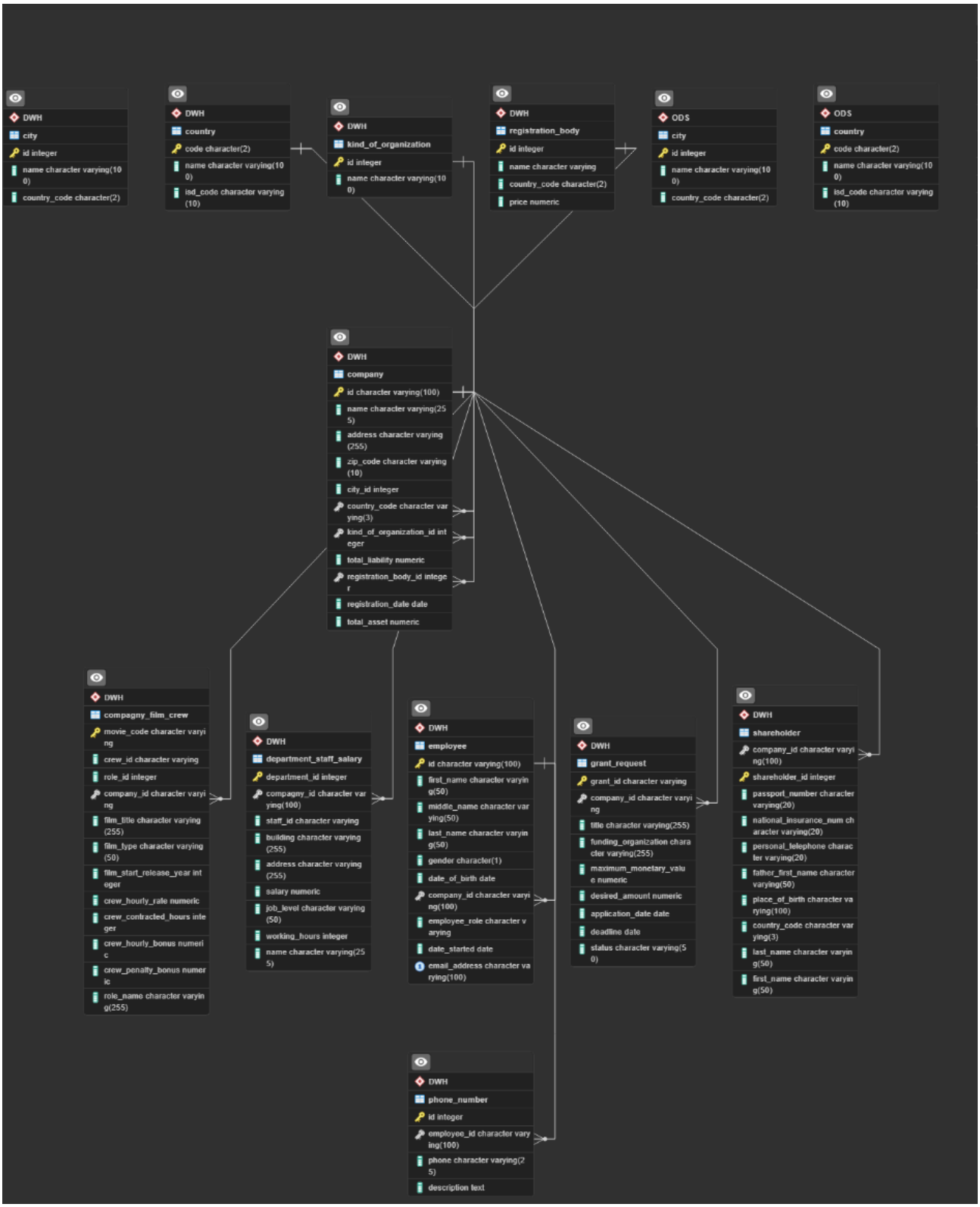


- Agrégation des tables (compagny_grant et grant_request) pour obtenir la table **grant_request**



E. Construction du Data Warehouse

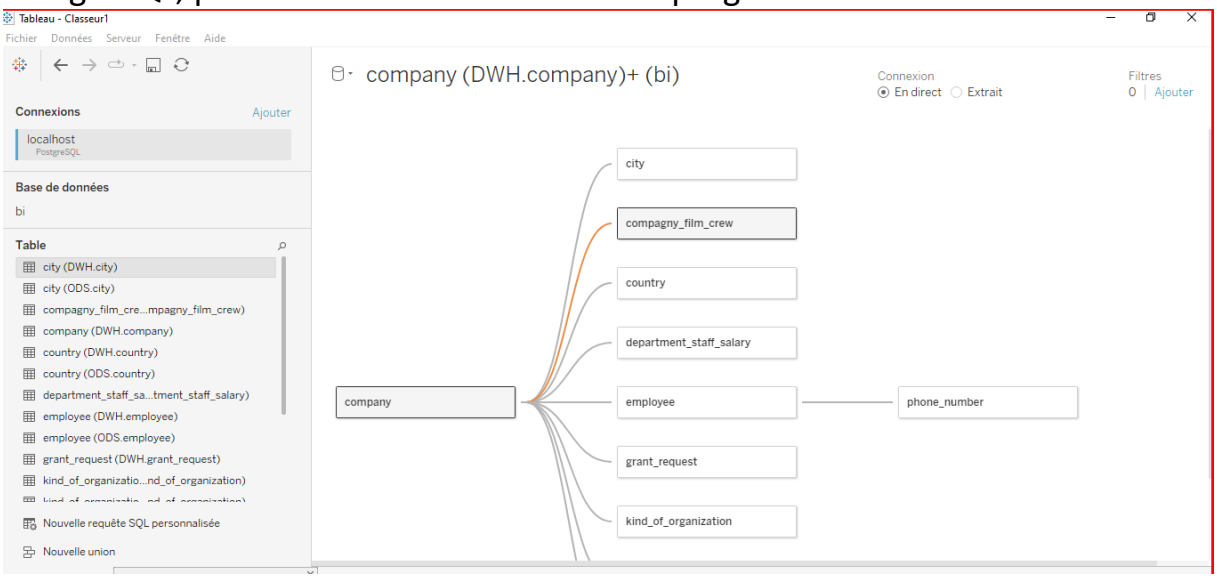
Nous avons le schéma final de notre data Warehouse qui est une réplique à l'identique de notre schéma de conception. En annexe vous trouverez le script SQL pour créer le même schéma.


DWH
city
id integer
name character varying(100)
country_code character(2)
DWH
country
code character(2)
name character varying(100)
isd_code character varying(10)
DWH
kind_of_organization
id integer
name character varying(100)
DWH
registration_body
id integer
name character varying
country_code character(2)
price numeric
ODS
city
id integer
name character varying(100)
country_code character(2)
ODS
country
code character(2)
name character varying(100)
isd_code character varying(10)
DWH
company
id character varying(100)
name character varying(255)
address character varying(255)
zip_code character varying(10)
city_id integer
country_code character varying(3)
kind_of_organization_id integer
total_liability numeric
registration_body_id integer
registration_date date
total_asset numeric
DWH
compagny_film_crew
movie_code character varying
crew_id character varying
role_id integer
company_id character varying
film_title character varying(255)
film_type character varying(50)
film_start_release_year integer
crew_hourly_rate numeric
crew_contracted_hours integer
crew_hourly_bonus numeric
crew_penalty_bonus numeric
role_name character varying(255)
DWH
department_staff_salary
department_id integer
compagny_id character varying(100)
staff_id character varying
building character varying(255)
address character varying(255)
salary numeric
job_level character varying(50)
working_hours integer
name character varying(255)
DWH
employee
id character varying(100)
first_name character varying(50)
middle_name character varying(50)
last_name character varying(50)
gender character(1)
date_of_birth date
company_id character varying(100)
employee_role character varying
date_started date
email_address character varying(100)
DWH
grant_request
grant_id character varying
company_id character varying
title character varying(255)
funding_organization character varying(255)
maximum_monetary_value numeric
desired_amount numeric
application_date date
deadline date
status character varying(50)
DWH
shareholder
company_id character varying(100)
shareholder_id integer
passport_number character varying(20)
national_insurance_num character varying(20)
personal_telephone character varying(20)
father_first_name character varying(50)
place_of_birth character varying(100)
country_code character varying(3)
last_name character varying(50)
first_name character varying(50)
DWH
phone_number
id integer
employee_id character varying(100)
phone character varying(25)
description text

IV. Visualisation avec Tableau

A. Connexion à la Base de Données

Instructions pour connecter Tableau à PostgreSQL : Installation de pilote PostgreSQL, puis connecter ODBC au fichier programmes de TABLEAU.

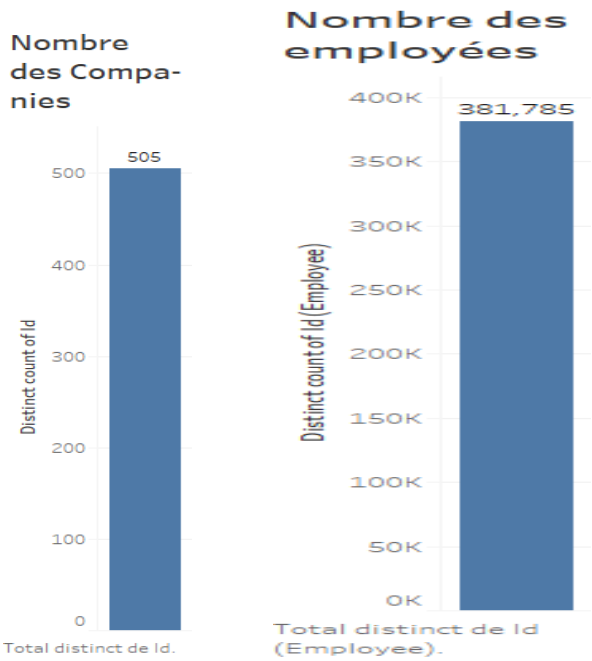


B. Création des visuels

Objectif de visualisation : créer un tableau de bord interactif qui offre une vue d'ensemble des relations et des détails au sein d'une organisation cinématographique.

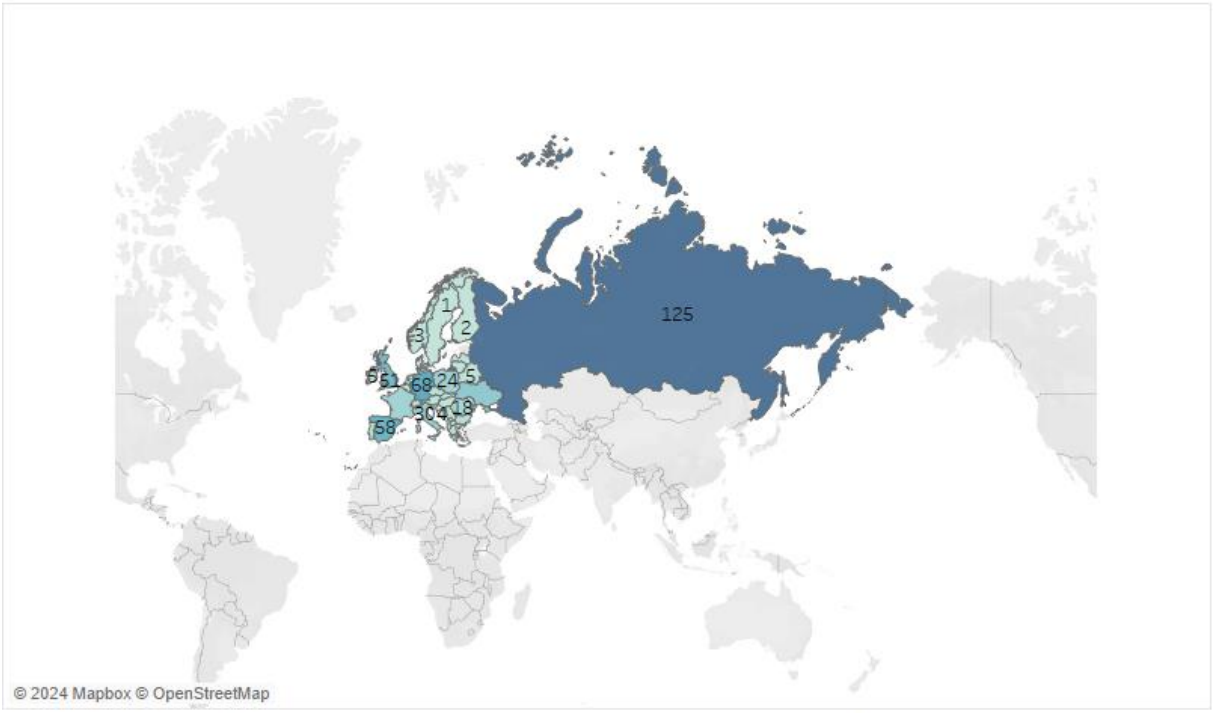
1. Vue d'ensemble de l'organisation :

- **Afficher le nombre total d'employés, de départements, de sociétés affiliées, de pays où l'organisation opère.**



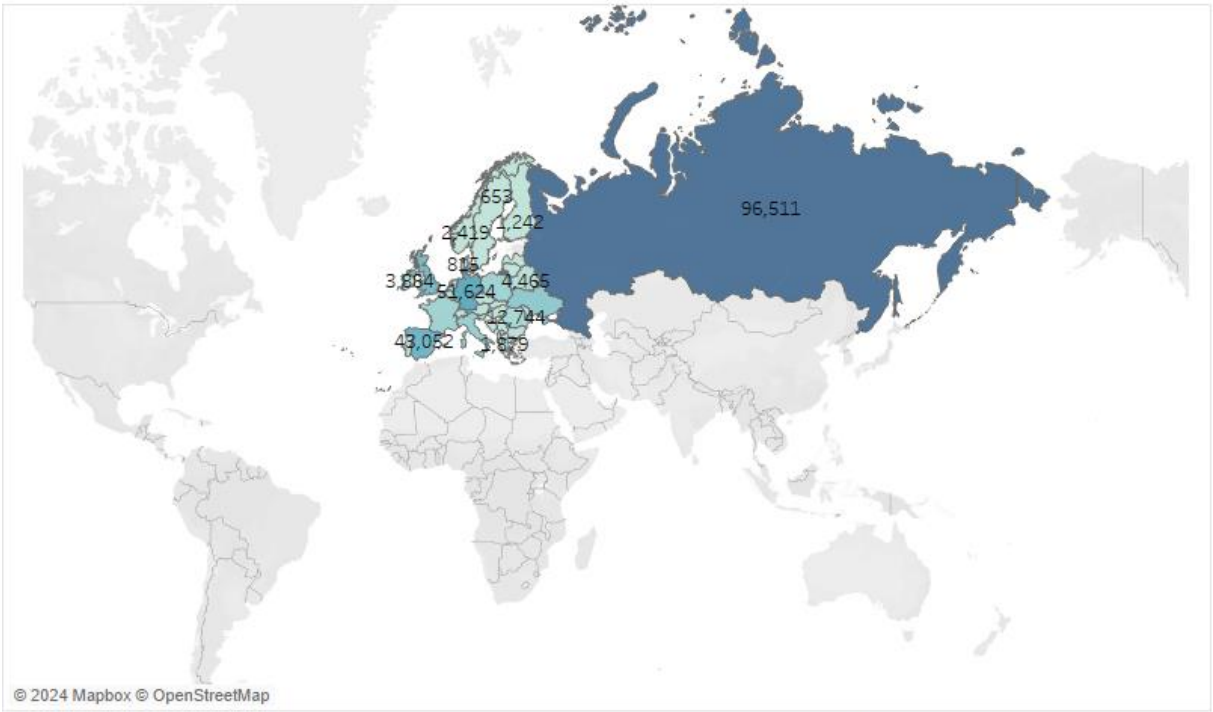
14

Repartition des Société par Pays



Carte basée sur les Longitude (générée) et Latitude (générée). La couleur met en avant le/la total distinct de Id. Les détails affichés sont associés au/à la Country Code.

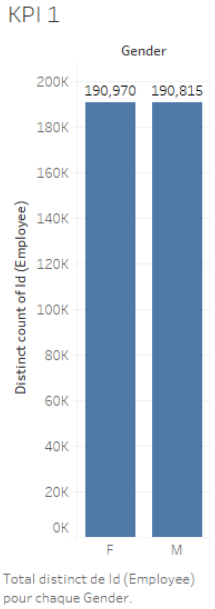
Nombre des employées par Pays



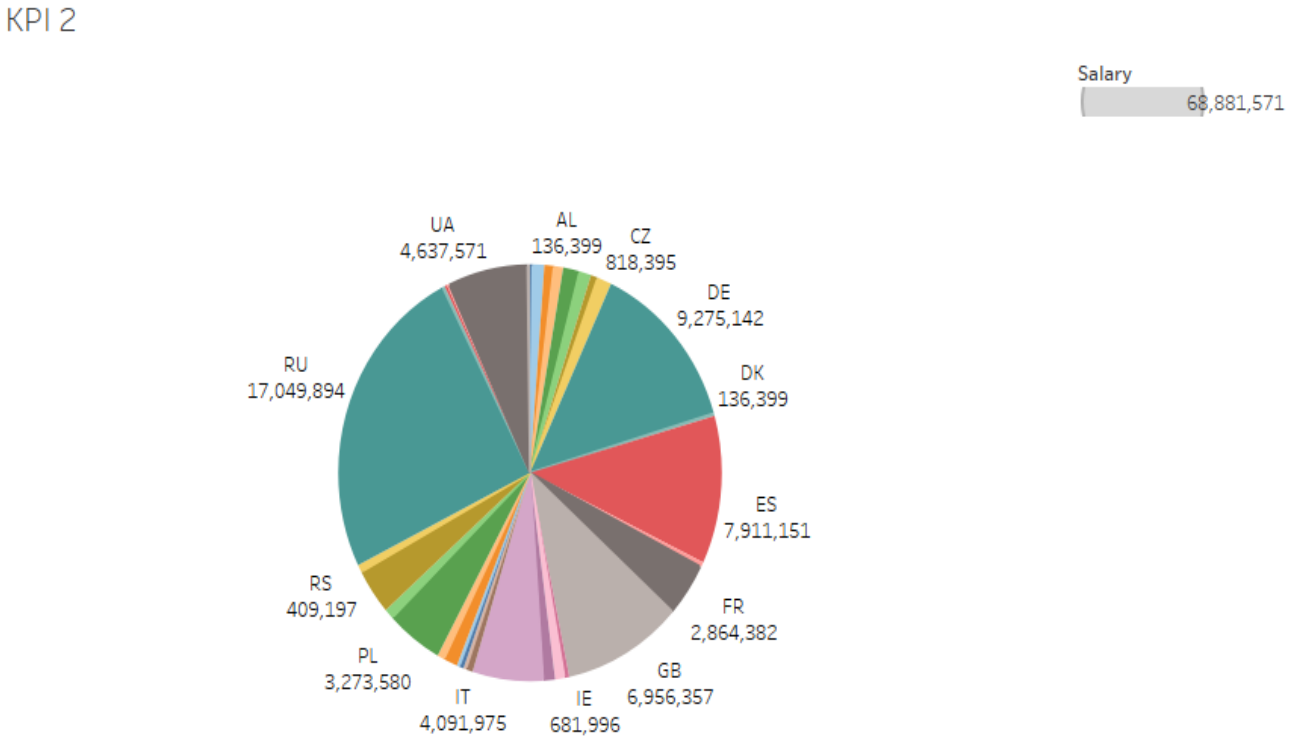
Carte basée sur les Longitude (générée) et Latitude (générée). La couleur met en avant le/la total distinct de Id (Employee). Les détails affichés sont associés au/à la Country Code.

2. Indicateurs KPI :

- **KPI1 : Ratio hommes-femmes**



- **KPI2 : Répartition des revenus par région**



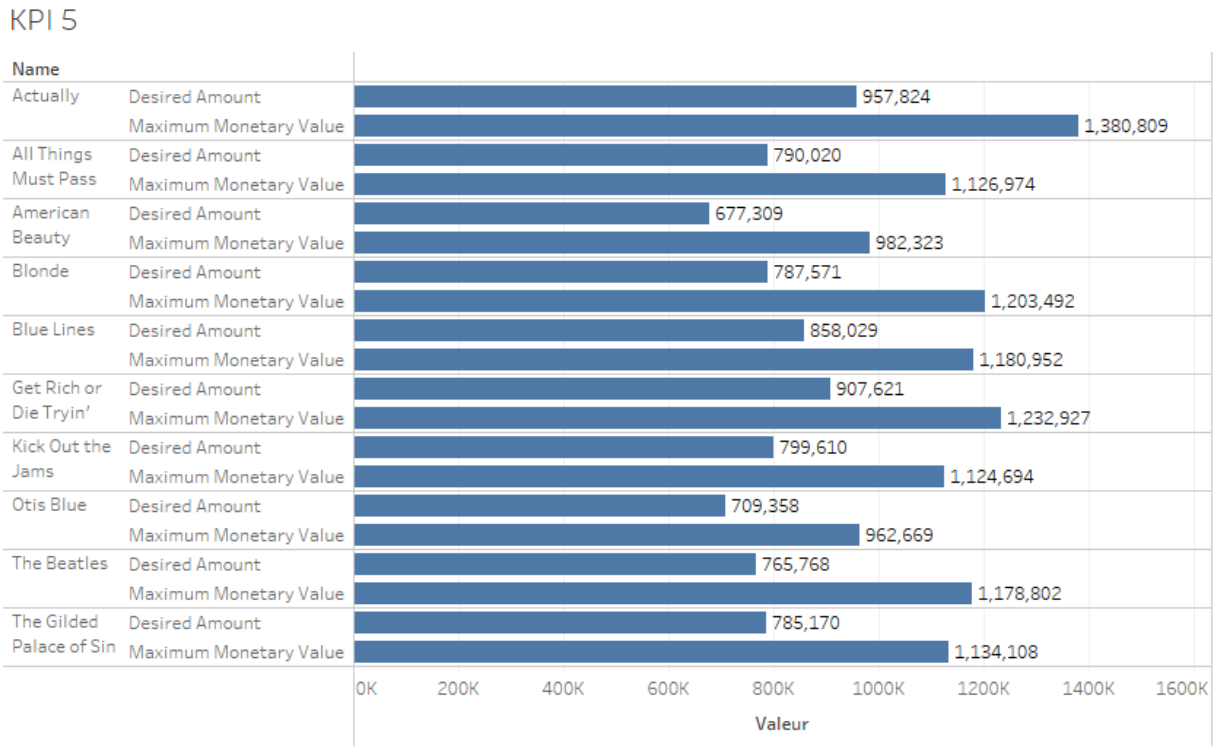
- **KPI3 : Nombre de prix remportés par Film**

Somme de Price pour chaque Name. La couleur affiche des détails associés au/à la Name. La vue est filtrée sur Name, qui conserve 10 membres sur 506.

- **KPI 4 : Nombre de prix remportées par Pays**

Name (Country) et somme de Price. La couleur met en avant le/la somme de Price. La taille correspond au/à la somme de Price. Les repères sont étiquetés par Name (Country) et somme de Price. La vue est filtrée sur Name (Country), qui exclut NULL.

- **KPI 5 : Evolution financier par film**



Desired Amount et Maximum Monetary Value pour chaque Name. La vue est filtrée sur Name, qui conserve 10 membres sur 505.

V. Conclusion

Ce projet de Business Intelligence a permis de mettre en place une solution complète pour l'analyse des données des compagnies de production cinématographique en Europe. En utilisant une combinaison de technologies avancées telles que PostgreSQL pour la gestion des bases de données, Talend pour les processus ETL, et Tableau pour la visualisation des données, nous avons pu créer un Data Warehouse robuste et performant.

Le processus a commencé par l'extraction des données provenant de différentes sources, notamment des fichiers CSV et des bases de données PostgreSQL. Ces données ont ensuite été transformées pour obtenir un modèle en étoile, facilitant ainsi les analyses rapides et efficaces. La phase de chargement a intégré ces données transformées dans notre Data Warehouse, prêt à être utilisé pour des visualisations interactives.

Les tableaux de bord créés offrent une vue d'ensemble détaillée des structures organisationnelles, des démographies et des performances financières des compagnies de production cinématographique. Les indicateurs clés de performance (KPI) tels que le ratio hommes-femmes, la répartition des revenus par région, et le nombre de prix remportés par film et par pays, ont permis d'obtenir des insights précieux pour la prise de décision stratégique.

En conclusion, ce projet démontre l'importance et l'efficacité d'une solution BI bien conçue pour le secteur cinématographique. Il fournit aux utilisateurs les outils nécessaires pour explorer et comprendre en profondeur les dynamiques internes et les performances de l'industrie, facilitant ainsi une prise de décision informée et stratégique. Les technologies utilisées ont prouvé leur capacité à gérer de grandes quantités de données et à offrir des visualisations claires et utiles, soulignant l'importance d'une intégration harmonieuse des différentes étapes de l'ETL et de la visualisation pour le succès d'un projet BI.